Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2018»

Москва, 30 мая — 2 июня 2018 г.

ОПЫТ ОБЪЕКТИВНОЙ ОЦЕНКИ ИНТОНАЦИОННОГО КАЧЕСТВА СИНТЕЗИРОВАННОЙ РУССКОЙ РЕЧИ

Лобанов Б. М. (Lobanov@newman.bas-net.by),
Соломенник А. И. (anna.i.prodan@gmail.com),
Житко В. А. (zhitko.vladimir@gmail.com)

Объединённый институт проблем информатики
НАН Беларуси, Минск, Беларусь

AN EXPERIENCE OF THE OBJECTIVE ESTIMATION OF INTONATION QUALITY OF THE SYNTHESIZED RUSSIAN SPEECH

Lobanov B. M. (Lobanov@newman.bas-net.by),
Solomennik A. I. (anna.i.prodan@gmail.com),
Zhitko V. A. (zhitko.vladimir@gmail.com)

United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The paper describes an experiment on an instrumental evaluation of the intonation quality of synthesized Russian speech by using of “Inton@Trainer” computer system. The system was originally designed to train learners in producing the basic intonation patterns of Russian speech. It is based on comparing the melodic portraits of a reference sentence and a sentence pronounced by the learner. Our approach to assessing the intonational quality of speech allows to treat a synthesized speech with the same strict requirements as are applied to students studying Russian as a second language. We describe the technology used for the instrumental evaluation of the intonation quality of synthesized speech and the acoustic database of reference phrases used to assess the intonation quality of synthesized

speech. The paper presents the results of testing the intonation quality of two Russian synthetic voices. We discuss the results of the experiment and outline the ways for improving the methods for objective evaluation of synthesized speech prosodic quality, as well as the possibility of applying the developed system in other linguistic tasks.

Keywords: Speech intonation, speech synthesis, objective evaluation, intonation quality, systems of analysis and assessment of intonation, learning system, Russian intonation

Введение

В работе [Lobanov, 2017] дано описание методики использования разработанной ранее компьютерной системы *"Inton@Trainer"* в качестве тренажёра на начальном этапе обучения РКИ в рамках освоения учащимися интонационных конструкций русской речи [Bryzgunova, 1968], [Odintsova, 2011]. Данная работа посвящена описанию эксперимента по нетрадиционному использованию разработанной компьютерной системы *"Inton@Trainer"*[1], а именно в задаче объективной (инструментальной) оценки интонационного качества синтезированной русской речи.

Как известно, существуют две основных разновидности методов оценки качества синтезированной речи: субъективный и объективный (инструментальный). Субъективный метод основываются на статистической обработке субъективных оценок качества синтезированной речи достаточно большим числом слушателей-экспертов.

Для русского языка настоящему времени достаточно разработанным методом оценки интонационного качества синтезированной речи является именно субъективный метод. В работе [Solomennik, 2013] описана методика и результаты субъективного тестирования интонационного качества синтезированной речи отечественных и зарубежных русскоязычных синтезаторов. Для прослушивания аудиторам были предложены 40 тестовых фраз, которые были синтезированы несколькими русскоязычными голосами разных синтезаторов.

В результате общая естественность интонации синтезированных голосов, участвовавших в тестировании, нормированная на естественный голос, изменяется от 49 % для голоса «Olga» (Loquendo TTS) до 72 % «Владимир» (VitalVoice TTS). Показатели, подсчитанные отдельно для каждого интонационного типа, позволили также найти слабые места каждой системы синтеза в отношении определённых интонационных конструкций. Самый низкий показатель точности был получен для вопросов с вопросительными словами и восклицаний. Длительный опыт работы авторов с синтезированной речью и проведённый эксперимент по влиянию различных ошибок на восприятие слушающими качества синтезированной речи [Solomennik, 2015] показывает, что неадекватное интонирование является основной проблемой качества современных русскоязычных синтезаторов.

---

[1] Больше информации о системе и бесплатное скачивание её демонстрационных версий доступно на сайте: https://intontrainer.by/

Проведение подобных тестов субъективными методами является довольно трудоёмкой задачей, и для того, чтобы ускорить процесс оценки, используются различные инструментальные методы, основывающиеся на автоматическом сравнении синтезированной и естественной речи [Falk, 2008]. Известны также исследования по оценке качества синтезированной речи с использованием систем автоматического распознавания речи [Bachan, 2012]. К объективным методам оценки можно отнести и систему [Norrenbrock, 2012], позволяющую инструментально оценивать качество просодической обработки.

Объективные методы основываются на оценке меры сходства интонации синтезированной и естественной речи с использованием специально разработанных инструментальных средств. В данной работе в качестве такого средства предлагается использовать систему *"Inton@Trainer"*, а в качестве эталонного тестового материала использовать ту же фразовую БД [Lobanov, 2017], которая была рекомендована И. В. Одинцовой [Odintsova, 2011] в качестве образцов для освоения учащимися интонационных конструкций русской речи.

Такой подход к оценке интонационного качества речи позволяет отнестись к синтезированной речи со столь же строгими требованиями, как и к учащимся на курсах РКИ.

1. Метод объективной оценки интонационного качества синтезированной речи

В основу методологии оценки интонационного качества положена количественная оценка различных мер сходства для интонационных параметров синтезированной и естественной речи, вычисляемая с помощью системы *"Inton@Trainer"*. Сходство определяется в рамках предложенной ранее [Lobanov, 2014)] модели представления интонационных конструкций русской речи (ИК-1 — ИК-7) — в виде Универсальных (Унифицированных) Мелодических Портретов (УМП) в нормированных координатах «Частота — Время». В качестве критериев сходства интонации синтезированной и естественной речи выступают степень их близости по диапазону изменения частоты основного тона (F0) и форме кривой, отображаемой в виде УМП. Вычисление этих критериев сходста, осуществляемое системой *"Inton@Trainer"*, наглядно отображается на экране. Пример экранного отображения результатов сравнения эталонной интонации (красная линия) с синтезированной (коричневая линия) для фразы «*Она уезжаент за+втра*.» представлен на Рис. 1.

На рисунке красный столбец слева отображает диапазон изменения F0 эталонной естественной фразы, а коричневый — синтезированной. Справа красным цветом отображается линейный график УМП эталонной естественной фразы, а коричневым — синтезированной фразы. Внизу под графиками приведены минимальное и максимальное значения F0 эталонной (Template) и синтезированной (User) фраз. Оценка интонационного сходства синтезированной и естественной речи осуществляется по двум основным критериям: а) по степени сходства диапазонов изменения F0 и б) по степени сходства УМП кривых.

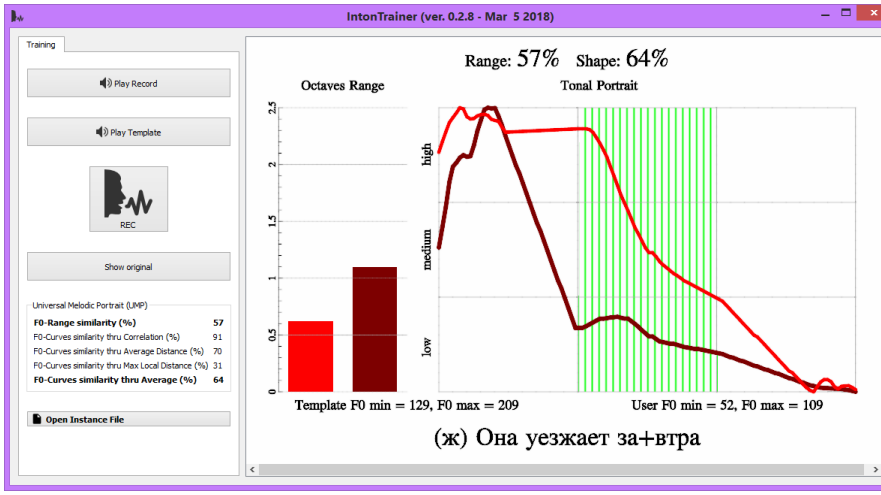

Рис. 1. Результат сравнения интонации фразы «*Она уезжаент за+втра.*» эталонной и синтезированной речи (Голос: Юрий — Nuance Vocalizer)

Вверху над графиками Рис. 1 приведены значения для мер сходства по диапазону — **Range 57%** и по форме кривой мелодического портрета — **Shape 64%**, вычисленные следующим образом.

***А) Оценка сходства диапазонов изменения F0* (Range Similarity).**
Определяется путём вычисления процентных соотношений — F0max / F0min между эталонным (templ) и синтезированным образцом (user) по формуле:

Range Similarity = {(F0max / F0min) user * 100%} / {(F0max / F0min) templ} (1)

***Б) Оценки сходства формы кривых УМП* (Shape Similarity).**
Вычисление **Shape Similarity** осуществляется следующими тремя различными способами.

- ***Корреляционная оценка сходства УМП***. Определяется путём расчёта взаимной корреляции — **r** (см.: http://statistica.ru/) между УМП эталонной и синтезированной фраз. В этом случае мера сходства — **Shape Similarity** определяется по формуле:

  Shape Similarity (thru correlation) = [(r+1) / 2] *100% (2)

- ***Среднее значение сходства УМП***. Определяется путём расчёта среднего значения векторного расстояния **d** между УМП эталонной и синтезированной фраз — Average Distance,

  Shape Similarity (thru average distance) = (1 — D)*100% (3)

- ***Минимальное значение сходства УМП***. Определяется путём расчёта максимального значения локального расстояния — Maximum Local Distance.

  Shape Similarity (thru max local distance) = (1 — dmax)*100% (4)

Использование 3-х различных мер сходства УМП позволяет оценить различие кривых не только в статистическом плане, но также в среднем и на тех участках, где имеется наибольшее отклонение от эталонной кривой.

Таким образом, оценка интонационного качества синтезированной речи, в сравнении с естественной, осуществляется по следующим четырём критериям:

1. Оценка сходства диапазонов изменения F0
2. Корреляционная оценка сходства УМП
3. Среднее значение сходства УМП
4. Минимальное значение сходства УМП

Подробные сведения, касающиеся технологии оценки интонационного качества, приведены на сайте https://intontrainer.by/. (См.: User Guide — Russian: «Анализатор и тренажёр речевой интонации»).

2. Результаты тестирования интонационного качества синтезированной речи

В качестве акустической базы данных эталонных ИК для сравнения с синтезированными фразами были использованы образцы фраз, произносимых диктором с женским голосом, содержащихся в мультимедийном учебнике РКИ [Odintsova, 2011]. Были отобраны 14 простых фраз по две на каждую ИК (см. Табл. 1) таким образом, чтобы по возможности исключить различные варианты прочтения (положение ядра и тип ИК).

Таблица 1. Тексты фраз

ИК1.	*Он гуляет.*	*Она уезжает завтра.*
ИК2.	*Какой?*	*Почему ты опоздал?*
ИК3.	*Любит?*	*Он завтра уезжает?*
ИК4.	*А дедушка?*	*А сегодня?*
ИК5.	*Кому она только не писала!*	*Какой сегодня день!*
ИК6.	*Весело как!*	*Какой фильм!*
ИК7.	*Да какая она актриса!*	*Да какая там выставка!*

Для тестирования были выбраны два наиболее качественных русскоязычных синтезаторов речи с синтезированными мужскими голосами:

- Владимир — VitalVoice TTS (ЦРТ, Россия).
- Юрий — Nuance Vocalize (Nuance Communications, Inc., США).
- Женский голос (Естественная речь).

Женский голос использовался для сравнения интонационного качества синтезированной и естественной речи. Фразы с ИК1 — ИК7 были начитаны диктором-женщиной без предварительного указания требуемой типа ИК.

Для всех вариантов тестирования в качестве эталонных использовались образцы фраз, произносимых диктором с женским голосом, взятые из мультимедийного учебника РКИ.

В таблицах 2, 3 приведены численные значения оценок интонационного сходства синтезированной речи для каждой из 7 интонационных конструкций по четырём вышеописанным критериям, а в таблице 4 — для женского голоса естественной речи.

Таблица 2. Оценки интонационного сходства для синтезатора — VitalVoice

Голос: Владимир				
Тип ИК	Диапазонная (%)	Корреляционная (%)	Средняя (%)	Минимальная (%)
ИК-1	22,5	68,5	63,5	32,5
ИК-2	79,5	66,0	65,0	31,5
ИК-3	30,5	78,0	66,0	32,0
ИК-4	36,5	45,0	54,5	19,5
ИК-5	47,5	78,5	72,0	31,0
ИК-6	43,0	53,5	57,5	36,5
ИК-7	23,5	59,5	57,5	22,5
Ср. знач.	40,4	64,1	62,3	29,4

Таблица 3. Оценки интонационного сходства для синтезатора — Nuance Vocalizer

Голос: Юрий				
Тип ИК	Диапазонная (%)	Корреляционная (%)	Средняя (%)	Минимальная (%)
ИК-1	35,5	89,0	75,5	34,0
ИК-2	68,0	82,0	73,5	50,0
ИК-3	31,5	90,5	83,5	60,5
ИК-4	54,5	30,0	47,0	13,5
ИК-5	45,5	90,5	78,0	32,0
ИК-6	45,0	50,5	55,0	14,0
ИК-7	13,5	58,0	58,0	33,0
Ср. знач.	41,9	70,1	67,2	33,9

Таблица 4. Оценки интонационного сходства для естественной речи

Женский голос				
Тип ИК	Диапазонная (%)	Корреляционное (%)	Средняя (%)	Минимальная (%)
ИК-1	73,0	96,0	79,0	52,5
ИК-2	79,0	82,5	72,5	41,5
ИК-3	22,0	88,5	73,5	39,5
ИК-4	53,5	76,0	69,0	20,0
ИК-5	53,5	95,0	83,5	58,0

Женский голос				
Тип ИК	Диапазонная (%)	Корреляционное (%)	Средняя (%)	Минимальная (%)
ИК-6	61,5	93,0	80,0	49,0
ИК-7	22,0	62,5	56,5	13,0
Ср. знач.	52,1	84,8	73,4	39,1

На рисунке 2 представлены графики усреднённых по ИК1 — ИК7 значений мер сходства с использованием каждого из 4-х видов оценки сходства: 1 — Диапазонной, 2 — Корреляционной, 3- Средней, 4 — Минимальной. Как видно из рисунка, каждый из видов оценки отображает одну и ту же тенденцию увеличения степени сходства в последовательности: Владимир (VitalVoice TTS); Юрий (Nuance Vocalizer); Женский голос (Естественная речь). В данной ситуации трудно отдать предпочтение какому-либо одному из видов оценки сходства. Это решение может быть сделано только в результате более масштабных исследований.

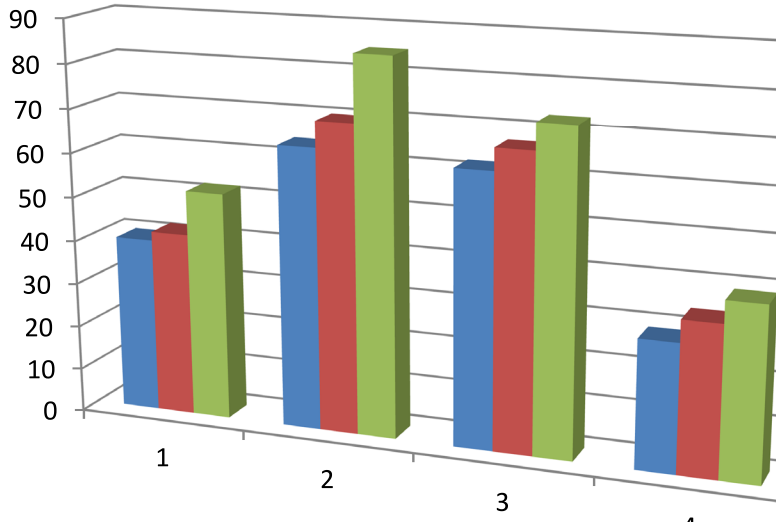

Рис. 2. Усреднённые по ИК1 — ИК7 значения мер сходства для каждого из 4-х видов оценок (Ряд 1 — VitalVoice TTS; Ряд 2 — Nuance Vocalizer; Ряд 3 — Ест. речь)

На рисунке 3 представлены графики усреднённых значений 4-х оценок сходства для каждой из ИК. Как видно из рисунка, тенденция, замеченная при рассмотрении Рис. 2 в основном сохраняется в большинстве случаев. Исключением являются ИК4 и ИК6, когда синтезатор «VitalVoice TTS» показывает лучшие результаты, чем «Nuance Vocalizer». Из рисунка видно также, что преимущество интонационного качества естественной речи, в значительной степени, проявляются в ИК4, ИК6, ИК7 и, в меньшей степени, в ИК2 и ИК3.

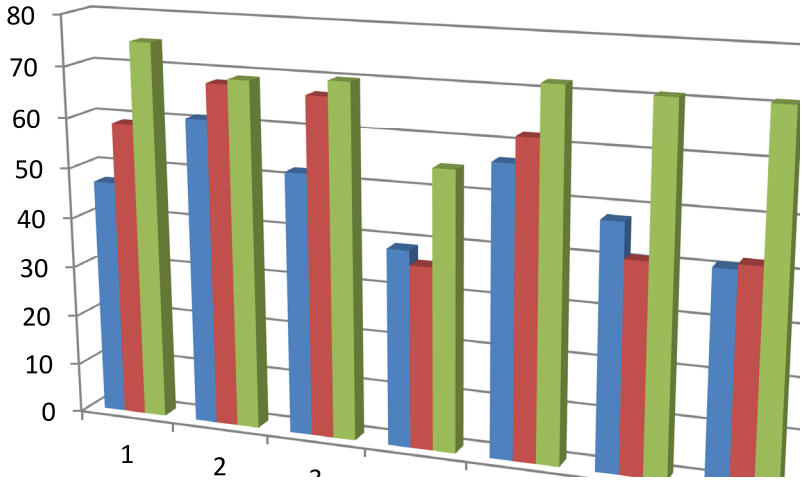

Рис. 3. Усреднённые значения 4-х оценок мер сходства для каждой из ИК (Ряд 1 — VitalVoice TTS; Ряд 2 — Nuance Vocalizer; Ряд 3 — Ест. речь)

Описанные результаты тестирования относятся к простейшим односинтагменным, одноакцентным фразам, которые в реальных текстах представлены лишь незначительным процентом. Практический интерес, однако, в наибольшей степени представляют сложные многосинтагменные фразы, в которых каждая из синтагм может иметь более одного акцента. Именно такого рода фразы повсеместно встречаются в деловых и художественных текстах, лежащих в основе создания звукового сопровождения ТВ-передач и синтезированных аудио книг.

Ниже будут представлены результаты сравнительной оценки интонационного качества естественной и синтезированной речи на примере сложной, многосинтагменной фразы: «Не успела за Андреем затвориться дверь, как я увидел в своем кабинете высокого, широкоплечего мужчину, державшего в одной руке бумажный сверток, а в другой — фуражку с кокардой». В качестве эталона используется соответствующий отрезок сигнала из аудиокниги по повести А. П. Чехова «Драма на охоте» в исполнении профессионального диктора Александра Балакирева. Аудирование этого отрезка показало, что он состоит из 5-ти 3-х акцентных синтагм. Для тестирования были выбраны те же русскоязычные синтезаторы речи с синтезированными женскими и мужскими голосами и естественная речь непрофессионального диктора:

- «Анна» и «Владимир» — VitalVoice TTS (ЦРТ, Россия).
- «Катя» и «Юрий» — Nuance Vocalize (Nuance Communications, Inc., США).
- «Борис» (Естественная речь непрофессионального диктора).

Усреднённые значения оценок интонационных мер сходства с эталонами для каждой из синтагм представлены в таблице 5.

Таблица 5. Усреднённые значения оценок интонационных мер сходства

Текст 5-ти синтагменной фразы: *«Не успела за Андреем затвориться дверь, как я увидел в своем кабинете высокого, широкоплечего мужчину, державшего в одной руке бумажный сверток, а в другой — фуражку с кокардой»*		**Марка синтезатора**				**Диктор**
		Vital Voice TTS		**Nuance Vocalizer TTS**		
		Анна	**Владимир**	**Катя**	**Юрий**	**Борис**
1	*Не успе+ла за Андре+ем затвориться две+рь,*	32	48	74	56	78
2	*как я уви+дел в свое+м кабине+те*	43	41	59	37	81
3	*высо+кого, широкопле+чего мужчи+ну,*	32	36	66	73	81
4	*держа+вшего в одной руке бума+жный све+рток,*	23	38	35	44	78
5	*а в друго+й — фура+жку с кока+рдой.*	57	21	38	76	89
Среднее значение для 5-ти синтагм		**39**	**37**	**54**	**57**	**81**

На рисунке 4 представлены средние значения оценок для 5-ти синтагм по каждому из видов синтезированной и естественной речи.

На рисунках 5 и 6 представлены графические результаты сравнения эталонной интонации (красная линия) фразы с естественной («Борис») и синтезированной речью («Владимир» — VitalVoice TTS).

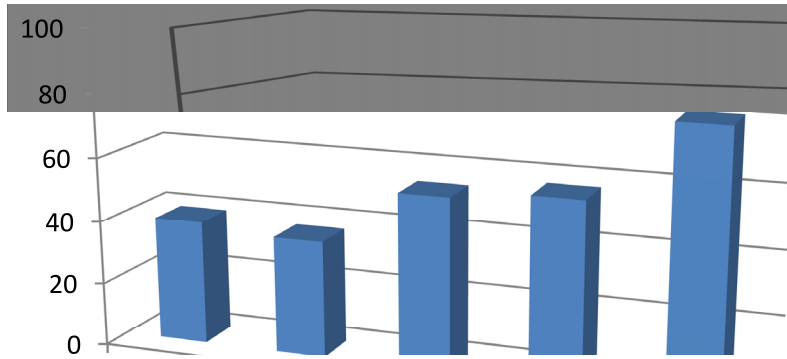

Рис. 4. Средние значения оценок для 5-ти синтагм (Ряд 1, 2 — VitalVoice TTS; Ряд 3, 4 — Nuance Vocalizer; Ряд 5 — Ест. речь)

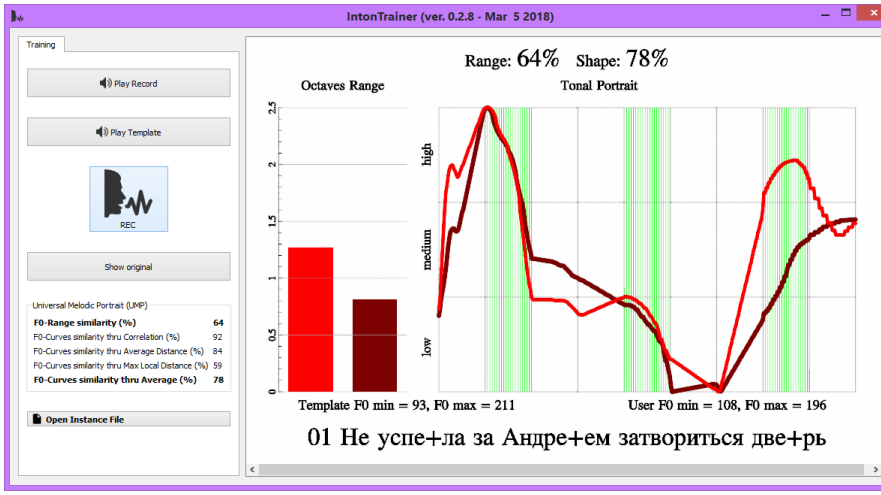

Рис. 5. Результат сравнения эталонной интонации (красная линия) с естественной речью («Борис»)

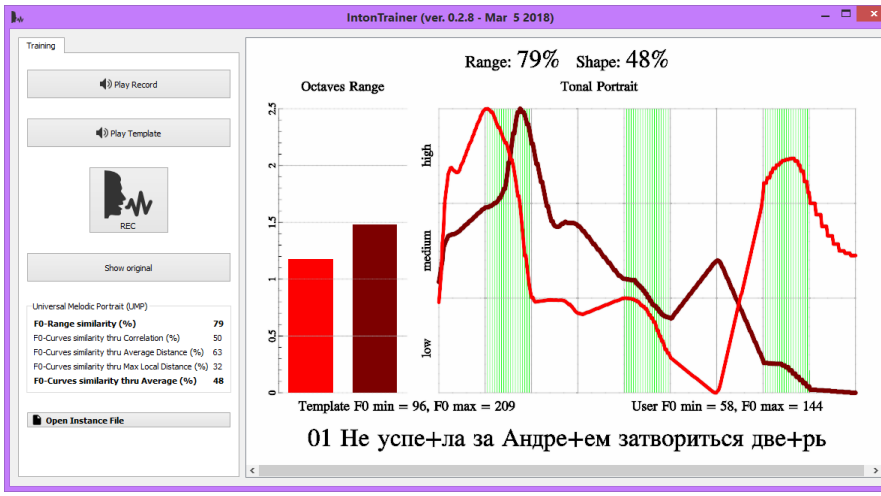

Рис. 6. Результат сравнения эталонной интонации (красная линия) с синтезированной речью («Владимир» — VitalVoice TTS)

Заключение

Представленные в данном докладе результаты объективной оценки интонационного качества синтезированной речи ни в коем случае не следует считать исчерпывающими. Это касается как слишком малого экспериментального материала, так и ограниченного числа исследуемых интонационных параметров. В частности, совершенно отсутствует информация о способах оценки энергетической, ритмической и паузальной компонент просодии речи.

Основным результатом этой работы мы считаем тот факт, что она продемонстрировала возможность использования разработанной системы *“Inton@ Trainer”* не только как нового компьютерного средства при обучении интонации в рамках РКИ, но и как нового компьютерного средства для интонационных исследований.

Литература

1. *Bachan J., Kuczmarski T., Francuzik P.* (2012), Evaluation of synthetic speech using automatic speech recognition, Proc. XIV International PhD Workshop OWD, Wisla, pp. 500–505.
2. *Bryzgunova E. A.* (1968), Sounds and Intonation of Russian Speech [Zvuki I Intonatsiya Russkoy Rechi], Nauka, Moscow.
3. *Falk T. H., Möller S.* (2008), Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems, IEEE Signal Processing Letters, Vol. 15, pp. 781–784.
4. *Lobanov B. M., Okrut T.* (2014), Universal Melodic Portraits of Intonation Patterns in Russian Speech [Universalnye melodicheskie portrety intonacionnyh konstrukciy russkoy rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2014” [Komp’yuternaya lingvistika i intellektual’nye tekhnologii: Materialy Mezhdunarodnoy konferentsii “Dialog-2014”], Bekasovo, pp. 330–339.
5. *Lobanov B. M., Zhitko V. A., Kharlamov A. A.* (2017), A computer system of teaching intonation patterns of Russian speech [Komp’yuternaya sistema obucheniya intonatsionnym konstruktsiyam russkoy rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp’yuternaya lingvistika i intellektual’nye tekhnologii: Materialy Mezhdunarodnoy konferentsii «Dialog-2017], Moscow, pp. 287–302.
6. *Norrenbrock C. R., Hinterleitner F., Heute U., Möller S.* (2012), Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals, Signal Processing Letters, IEEE, pp. 255–258.
7. *Odintsova I. V.* (2011), Sounds. Rhythm. Intonation. [Zvuki. Ritm. Intonatsiya.] — Flinta-Nauka, Moscow.
8. *Solomennik A. I., Cherentsova A. E.* (2013), A Method for Auditory Evaluation of Synthesized Speech Intonation, Miloš Železný et al. (Eds.): SPECOM 2013, Lecture Notes in Artificial Intelligence 8113, Springer, pp. 9–16.
9. *Solomennik A. I.* (2015) An influence of defects in synthesized speech on its naturalness [Zavisimost’ estestvennosti zvuchaniya sintezirovannoy rechi ot nalichiya oshibok razlichnykh tipov], Actual problems of philological science: the view of a new generation. Reports of participants of the XX-XXI International conferences of students, graduate students and young scientists «Lomonosov». Section «Philology» [Aktual’nye problemy filologicheskoy nauki: vzglyad novogo pokoleniya. Doklady uchastnikov XX-XXI Mezhdunarodnykh konferentsiy studentov, aspirantov i molodykh uchenykh “Lomonosov”. Sektsiya “Filologiya”], Moscow, pp. 475–480.